[SQUEAKING]

[RUSTLING]

[CLICKING]

YUFEI ZHAO: In this video, we'll look at three basic yet important inequalities in probability-- Markov's inequality, Chebyshev's inequality, and the Chernoff bound. Markov's inequality says that if we are given X, a real valued non-negative random variable, then for every positive number lambda we have the following inequality-- the probability that X is bigger-- is at least lambda, this probability is no more than the expectation of X divided by lambda.

One way to interpret this inequality is that if X is a non-negative random variable with small expectation, then it is unlikely for X to be very high. This is a very important and useful inequality. And let us prove it. The proof is quite short.

We can start with the expectation of X. And rewrite this in the following way. It is at least the expectation of X times the indicator function corresponding to when X is at least lambda. So this means it is 1 if X is at least lambda and 0 otherwise.

Now, claim next that the expression inside the expectation is at least lambda times 1 sub X being at least lambda, because when X is at least lambda, well, X is at least a lambda. Otherwise, both sides are 0. So this inequality holds as well.

And finally, we can pull out the lambda and get lambda times the probability that X is at least lambda. So this finishes the proof of Markov's inequality.

Next, let's move to Chebyshev's inequality. The statement of Chebyshev's inequality is that if we're given a real random variable X, then for every positive real number lambda, the probability that X deviates from its mean by at least lambda times the square root of the variance of X-- so this probability-- is at most 1 over lambda squared.

Let me remind you that the variance of the random variable X is the quantity which is defined to be the expectation of X minus the expectation of X squared. And it is also equal to the expectation of X squared minus the square of the expectation of X. So this is the variance of the random variable X.

Yeah. So intuitively, what Chebyshev's inequality tells us is that if a random variable has small variance, then it is unlikely to deviate too far from its mean. Let us now prove this inequality.

The left-hand side, namely this probability, we can rewrite it by squaring the inequality in this expression, which we-- doing this gets us the following. On the left-hand side, we get X minus the expectation of X squared. On the right-hand side, we get lambda times lambda squared times the variance of X.

And now let us apply Markov's inequality. So apply what we saw earlier, apply Markov, to get the expectation of X minus the expectation of X squared divided by lambda squared, variance of X. But you see, from the definition of variance, the expression circled in red here is equal to 1. And therefore, the right-hand side is equal to 1 over lambda squared. So that finishes the proof of Chebyshev's inequality.

The third inequality that we'll look at is known as the Chernoff bound. And unlike the earlier inequalities, which is for fairly general random variables, the Chernoff bound is for a more specific setting, namely a sum of independent random variables. And more specifically, let us look at the case where the random variable S sub n is a sum of n different plus minus 1's. So each Xi is plus 1 with probability 1/2 and minus 1 with probability 1/2, all chosen independently at random.

So one way to interpret Sn is that we walk on the number line, take n steps. Each step, we flip a coin and whether walk either 1 step to the right or 1 step to the left and walk n steps. The conclusion of Chernoff bound is that for every positive lambda, the probability that S sub n is at least lambda times root n is at most this quantity here, which is e to the minus lambda squared over 2.

So in other words, let S sub n-- so where we end up in this walk-- cannot drift too far away from the origin. And too far here means some large multiple of root n. It is worth noting that root n here is the variance of-- so the square root of the variance of S sub n.

So if we simply apply Chebyshev's inequality to S sub n, we would arrive at the following conclusion. So Chebyshev tells us that the probability that Sn is at least lambda is at most 1 over lambda squared, which is already an interesting bound. But the Chernoff bound gives a much stronger conclusion. So the right-hand side decreases very rapidly as a function of lambda compared to the Chebyshev bound, which decreases only at the rate of 1 over lambda squared.

Let us now prove Chebyshev-- the Chebyshev bound. For the proof, we'll need to introduce a new idea. And this is the idea of a moment-generating function. So let t be a non-negative real number. And let us consider the moment-generating function, in this case given as follows.

Consider the expectation of the quantity e to the t times S sub n. So instead of considering, for example, the expectation of S sub n, let's consider the exponential applied to the random variable. So this is the expectation is over the randomness in S sub n.

Well, let us now rewrite this expression by expanding the definition of S sub n. And we get the following. And the next step is where we crucially use the fact that the Xi's are independent random variables, which then allows us to split this expectation as a product of individual expectations.

Let's look at one of these terms, one of these expectation factors. Well, X sub 1 is plus 1 with probability 1/2 and minus 1 with probability 1/2. So this factor here equals to e to the t plus e to the minus t over 2. And likewise with all the other terms. So the right-hand side equals to this quantity raised to the power n.

So this exactly computes the moment-generating function of this random variable S sub n. Let us now try to manipulate this right-hand side expression to make it easier to work with. So starting with this expression, for now temporarily dropping the power of n, so just look at what's inside the parentheses, we can apply a Taylor series expansion.

So by Taylor series, I'm going to do the steps slightly on the quick side, but I encourage you to work it out yourself if you can follow this step. Basically, we apply the Taylor series expansion and note that the odd index terms all cancel each other out. So again, I encourage you to try out this calculation yourself. We get the following.

On the other hand, let's look at the expression e to the t squared over 2. So we'll see where this comes from in a second. So this X should be a t. So let's look at this e to the t squared over 2.

By using Taylor expansion over here, we get the following. X of each term is t to the 2k over k factorial, 2 to the k. So this is by writing out the Taylor series expansion for both of these expressions.

And finally, let's note that even on a term-by-term basis, one has this inequality by comparing what happens in the denominators. So I'll leave this as an exercise. But this is not too hard to see.

This allows us then to rewrite the previous line by upper bounding this moment-generating function by the expression e to the nt squared over 2. So that's what we just deduced.

Finally, let us use Markov's inequality to upper bound the probability that S sub n is at least lambda times root n. So that's what we're trying to bound. As with earlier, well, previously we squared both sides. But now let's take an exponential of both sides.

So for this step, actually, I think I will need t to be strictly positive for this step to hold. So let me change the hypothesis on t. So this is true. And then we can apply Markov to upper bound this probability by the expectation of this expression e to the t sub n, which is random, and then divided by the right-hand side, which is a constant.

And earlier, we had determined an upper bound for the moment-generating function. And plugging everything in, we get this following expression-- e to the nt squared over 2 minus t lambda square root n.

So this expression, this inequality, is true for every positive value of t. And the next step then is to use a value of t that works most in our favor. And one can do this by trying to minimize the expression on the right-hand side. And you can do this. And turns out that the optimal t to set is lambda over root n.

So you can determine the optimal value of t by taking the derivative. And this is the optimal value. And set t into this bound gives the desired conclusion.

So if you set this value of t into this right-hand side bound, you will see the right-hand side of Chernoff bound popping out. So that's the proof of the Chernoff bound. So the key idea introduced here is the moment-generating function. And because of the independence of the components of S sub n, we can factorize the moment-generating function and, through some additional algebraic manipulations, derive the desired bound.

So all three of these inequalities are basic and powerful techniques in probability. The Chernoff bound in particular gives an extraordinarily good bound on the probability that a sum of independent random variables can deviate from its expectation. And over here, although we only looked at the case when Xi is plus 1 or minus 1, each with probability 1/2, the same proof can be applied in a more versatile way to other settings, as long as you have a sum of independent random variables.